Decision Trees

David Suffolk

5/3/2020

Decision Trees

Import Data

```
library(rpart)
library(rpart.plot)
library(ggplot2)
data <- read.csv("./Project 02/nwCrow_bloodParasites_alaska_smith_2007_2008/nwCrow_sampling_alaska_smit
data_2 <- read.csv("./Project 02/nwCrow_bloodParasites_alaska_smith_2007_2008/nwCrow_bloodParasites_ala
de <- merge(data, data_2, by=0, all=TRUE)
head(de)
##
    Row.names Field.ID
                             DATE LOC
                                        LAT
                                                LONG SEX AGE AKD TARSUS WING MASS
## 1
                  75001 3/20/2007 SEWA 60.11 -149.44 1
                                                                   55.8 283
            1
                                                           1
## 2
           10
                  75010 3/22/2007 KENA 60.55 -151.23
                                                               0
                                                                   48.6 271
                                                                              390
                                                       2
                                                           1
## 3
           100
                  75100 3/12/2008 VALD 61.12 -146.35
                                                       2
                                                           1
                                                               0
                                                                   44.6
                                                                         264
                                                                              317
## 4
           101
                  86701 3/12/2008 VALD 61.12 -146.35
                                                       2 1
                                                               0
                                                                   47.1
                                                                         269
                                                                              343
                                                           2
## 5
           102
                  86702 3/12/2008 VALD 61.12 -146.35
                                                                   52.2
                                                                         291
                                                                              415
           103
                  86703 3/12/2008 VALD 61.12 -146.35
## 6
                                                           2
                                                               0
                                                                   47.0 266
                                                                              325
                                                       1
    Extraction.. LEUC1 LEUC2 HAEM1 HAEM2 PLAS1 PLAS2 Leuc_GenBank_Accession
##
         NOCRO01
                      0
## 1
                            0
                                  0
                                        0
                                              0
                                                    0
## 2
          NOCRO10
                      0
                            0
                                  0
                                        0
## 3
         NOCR100
                                  0
                                        0
                                              0
                                                    0
                      1
                            1
                                                                    MG765394
## 4
         NOCR101
                      0
                            0
                                  0
                                        0
                                  0
                                        0
                                                                    MG765394
## 5
         NOCR102
                      1
                                              0
                                                    0
## 6
                                  0
                                                                    MG765394
         NOCR103
                      1
                            1
                                        0
                                                    0
    Haem_GenBank_Accession Plas_GenBank_Accession
##
## 1
## 2
## 3
## 4
## 5
## 6
```

One Hot Encoding

```
for(unique_value in unique(de$LOC)){
de[paste("LOC", unique_value, sep = ".")] <- ifelse(de$LOC == unique_value, 1, 0)
}
head(de)
     Row.names Field.ID
                              DATE LOC
                                           LAT
                                                  LONG SEX AGE AKD TARSUS WING MASS
## 1
             1
                   75001 3/20/2007 SEWA 60.11 -149.44
                                                              1
                                                                       55.8
                                                                             283
                                                                                  448
                                                          1
                                                                  1
## 2
            10
                   75010 3/22/2007 KENA 60.55 -151.23
                                                          2
                                                              1
                                                                  0
                                                                       48.6
                                                                             271
                                                                                  390
                   75100 3/12/2008 VALD 61.12 -146.35
## 3
           100
                                                          2
                                                                       44.6
                                                                             264
                                                                                  317
                                                              1
                                                                  0
## 4
           101
                   86701 3/12/2008 VALD 61.12 -146.35
                                                          2
                                                              1
                                                                       47.1
                                                                             269
                                                                                  343
                                                                  0
## 5
                                                          2
                                                              2
           102
                   86702 3/12/2008 VALD 61.12 -146.35
                                                                       52.2
                                                                             291
                                                                                  415
## 6
           103
                   86703 3/12/2008 VALD 61.12 -146.35
                                                          1
                                                              2
                                                                  0
                                                                       47.0
                                                                             266
                                                                                  325
     Extraction.. LEUC1 LEUC2 HAEM1 HAEM2 PLAS1 PLAS2 Leuc_GenBank_Accession
##
## 1
          NOCRO01
                       0
                             0
                                    0
                                          0
                                                0
                                                       0
## 2
          NOCRO10
                       0
                             0
                                    0
                                          0
                                                0
                                                       0
## 3
          NOCR100
                                    0
                                          0
                                                                       MG765394
                       1
                             1
                                                0
                                                       0
## 4
          NOCR101
                       0
                             0
                                   0
                                          0
                                                0
## 5
          NOCR102
                       1
                             0
                                    0
                                          0
                                                0
                                                       0
                                                                       MG765394
## 6
          NOCR103
                       1
                             1
                                    0
                                          0
                                                                       MG765394
     Haem_GenBank_Accession Plas_GenBank_Accession LOC.SEWA LOC.KENA LOC.VALD
##
## 1
## 2
                                                             0
                                                                                0
                                                                       1
## 3
                                                             0
                                                                       0
                                                                                1
## 4
                                                             0
                                                                       0
                                                                                1
## 5
                                                             0
                                                                       0
                                                                                1
                                                                       0
## 6
                                                             0
                                                                                1
     LOC.HAIN LOC.JUNE LOC.HOME
## 1
            0
                      0
## 2
            0
                      0
                               0
## 3
            0
                      0
                               0
## 4
            0
                      0
                               0
## 5
            0
                      0
                               0
## 6
            0
                      0
                               0
```

Filter Columns

```
de <- de[,c(7,8,9,10,11,12,14,16,18,23,24,25,26,27,28)]
head(de)
```

```
SEX AGE AKD TARSUS WING MASS LEUC1 HAEM1 PLAS1 LOC.SEWA LOC.KENA LOC.VALD
##
## 1
       1
            1
                1
                    55.8
                           283
                                 448
                                          0
                                                0
                                                       0
                                                                 1
                                                                           0
                                                                                     0
## 2
       2
                0
                     48.6
                                 390
                                          0
                                                0
                                                       0
                                                                 0
                                                                           1
                                                                                     0
            1
                           271
## 3
       2
            1
                0
                     44.6
                           264
                                 317
                                                       0
                                                                 0
                                                                           0
                                                                                     1
## 4
       2
                0
                     47.1
                           269
                                 343
                                                0
                                                       0
                                                                 0
                                                                           0
            1
                                          0
                                                                                     1
## 5
       2
            2
                0
                    52.2
                           291
                                 415
                                          1
                                                0
                                                       0
                                                                 0
                                                                           0
                                                                                     1
## 6
            2
                0
                     47.0 266
                                 325
                                          1
                                                                           0
                                                                                     1
       1
     LOC.HAIN LOC.JUNE LOC.HOME
## 1
             0
                       0
                                 0
```

```
## 2
              0
## 3
       0
              Ω
                     0
## 4
       0
             0
                     0
## 5
       0
             0
                     0
        0
## 6
              0
                     0
```

Factoring

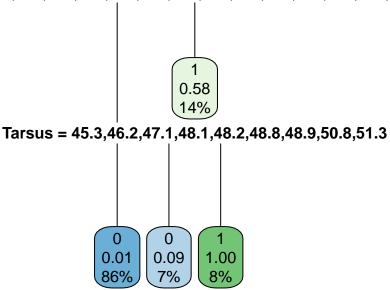
```
de$SEX <- as.factor(de$SEX)
de$AGE <- as.factor(de$AGE)
de$AKD <- as.factor(de$TARSUS)
de$TARSUS <- as.factor(de$TARSUS)
de$WING <- as.factor(de$WING)
de$MASS <- as.factor(de$MASS)
de$LEUC1 <- as.factor(de$LEUC1)
de$HAEM1 <- as.factor(de$HAEM1)
de$PLAS1 <- as.factor(de$PLAS1)
de$LOC.SEWA <- as.factor(de$LOC.SEWA)
de$LOC.KENA <- as.factor(de$LOC.KENA)
de$LOC.VALD <- as.factor(de$LOC.VALD)
de$LOC.HAIN <- as.factor(de$LOC.HAIN)
de$LOC.JUNE <- as.factor(de$LOC.JUNE)
de$LOC.HAIN <- as.factor(de$LOC.HAIN)</pre>
```

AKD Decision Tree

```
names(de) <- c("Sex","Age","AKD","Tarsus","Wing","Mass","LEUC1","HAEM1","PLAS1","SEWA","KENA","VALD","H
ran <- sample(1:nrow(de), 0.9 * nrow(de))
data_train <- de[ran,]
data_test <- de[-ran,]
dtm <- rpart(AKD~., data_train, method="class")
rpart.plot(dtm, compress=TRUE, uniform=TRUE)</pre>
```



i7,362,363,364,365,367,368,370,371,372,373,374,375,377,378,379,382,383,384,386,38



```
p <- predict(dtm, data_test, type="class")
confMat <- table(data_test$AKD,p)
accuracy <- sum(diag(confMat))/sum(confMat)
return (accuracy*100)</pre>
```

[1] 89.47368

LEUC1 Decision Tree

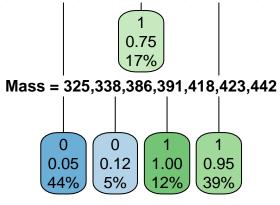
```
names(de) <- c("Sex","Age","AKD","Tarsus","Wing","Mass","LEUC1","HAEM1","PLAS1","SEWA","KENA","VALD","H
ran <- sample(1:nrow(de), 0.9 * nrow(de))
data_train <- de[ran,]
data_test <- de[-ran,]
dtm <- rpart(LEUC1~., data_train, method="class")
rpart.plot(dtm, compress=TRUE, uniform=TRUE)</pre>
```



349,351,353,355,356,358,359,362,365,367,368,372,373,374,375,384,386,387,388,391,



46.5,46.6,46.7,46.9,47.1,47.2,47.3,47.4,47.5,47.6,48.1,48.3,48.4,48.7,48.9,49.2,49.4,49.



```
p <- predict(dtm, data_test, type="class")
confMat <- table(data_test$LEUC1,p)
accuracy <- sum(diag(confMat))/sum(confMat)
return (accuracy*100)</pre>
```

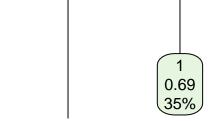
[1] 42.10526

HAEM1 Decision Tree

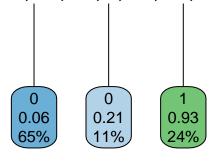
```
names(de) <- c("Sex","Age","AKD","Tarsus","Wing","Mass","LEUC1","HAEM1","PLAS1","SEWA","KENA","VALD","H
ran <- sample(1:nrow(de), 0.9 * nrow(de))
data_train <- de[ran,]
data_test <- de[-ran,]
dtm <- rpart(HAEM1~., data_train, method="class")
rpart.plot(dtm, compress=TRUE, uniform=TRUE)</pre>
```



i1,353,356,357,358,359,362,364,365,366,370,373,377,379,383,384,387,390,395,397,40



Tarsus = 44.1,45.4,46.3,46.5,47,47.1,47.6,48.1,48.7,48.8,49.3,49.5,52.2



```
p <- predict(dtm, data_test, type="class")
confMat <- table(data_test$HAEM1,p)
accuracy <- sum(diag(confMat))/sum(confMat)
return (accuracy*100)</pre>
```

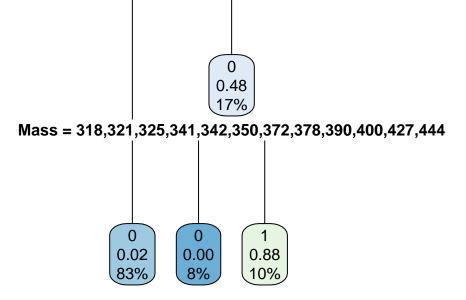
[1] 47.36842

PLAS1 Decision Tree

```
names(de) <- c("Sex","Age","AKD","Tarsus","Wing","Mass","LEUC1","HAEM1","PLAS1","SEWA","KENA","VALD","H
ran <- sample(1:nrow(de), 0.9 * nrow(de))
data_train <- de[ran,]
data_test <- de[-ran,]
dtm <- rpart(PLAS1~., data_train, method="class")
rpart.plot(dtm, compress=TRUE, uniform=TRUE)</pre>
```



47.2,47.3,47.5,47.6,47.7,48.1,48.2,48.3,48.4,48.5,48.7,48.8,48.9,49.2,49.3,49.5,49.6,49.



```
p <- predict(dtm, data_test, type="class")
confMat <- table(data_test$PLAS1,p)
accuracy <- sum(diag(confMat))/sum(confMat)
return (accuracy*100)</pre>
```

[1] 84.21053